DOI:10.1145/3512728

BY TOMMASO DI NOIA, NAVA TINTAREV, PANAGIOTA FATOUROU, AND MARKUS SCHEDL

Recommender Systems under European AI Regulations

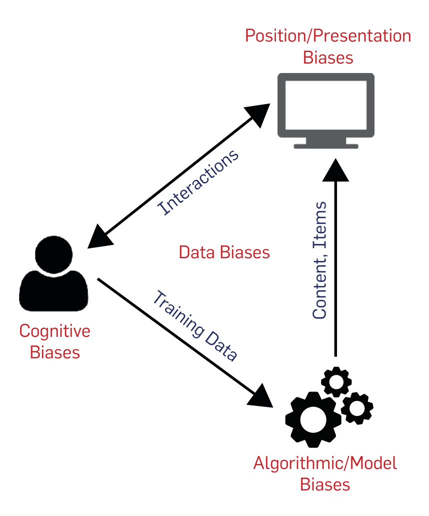
THE EUROPEAN COMMISSION (EC) has acknowledged the importance artificial intelligence (AI) plays in forming Europe's future, identifying AI as the most strategic technology of the 21st century.[a] With a recent proposal on a *Regulation Laying Down Harmonised Rules on Artificial Intelligence*[b] (EU Regulatory Framework for AI), the EC aims at introducing the first comprehensive legal framework on AI, which will identify specific risks for AI, provide a collection of high-risk application domains, propose specific requirements that AI systems should meet when used in such domains, and define obligations for users and providers (U.S. regulatory development relating to AI[c]). What clearly emerges from these efforts is the need for an AI that behaves in a responsible way. A clear and globally accepted definition of responsibility for AI systems is still under development, but will likely include notions such as fairness, security and privacy, explainability, safety, and reproducibility. Although safety and reproducibility are fundamental issues in AI research and its industrial application, we will not cover them here since they are requirements in many areas of technology, therefore not specific to AI.

According to the EC regulation, AI should be used in compliance with the E.U. Charter of Fundamental Rights,[d] including the right not to be discriminated against, the respect for private life, and the protection of personal data. The regulation also stresses the "*obligations for ex ante testing, risk management and human oversight of AI systems to minimize the risk of erroneous or biased AI-assisted decisions in critical areas such as education and training, employment, important services, law enforcement and the judiciary*." High-risk AI systems should meet specific legal requirements in relation to data management, documentation, human oversight, transparency, robustness, accuracy, and security. According to Article 10, "*training, validation and*

[a] See https://bit.ly/3HTQMP3
[b] See https://bit.ly/34vooEz
[c] See https://bit.ly/3rc2DkO
[d] See https://bit.ly/3r3mrH8

While much research has been devoted to uncover and mitigate biases in recommender systems, many research gaps still exist.

Figure 1. Different categories of biases and their interplay.



testing data sets shall be subject to appropriate data governance and management practices" which shall concern, in particular, "*examination in view of possible biases*" and "*identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed*." On the other hand, Article 15 is devoted to accuracy, robustness, and cybersecurity: high-risk AI systems must achieve all three throughout their entire life cycle to a satisfactory degree based on state-of-the-art security and privacy-preserving measures. The regulation makes it also clear that "*AI systems should be sufficiently transparent, explainable and well-documented*" (Article 13).

In the following, we attempt to provide the European scene for fairness, security and privacy, and explainability under the lens of recommender systems (RSs). Given their user-centric nature, RSs are fully touched by the principles and rules stated in the aforementioned EC documents, and therefore represent an interesting workbench to study their application. Issues related to fairness, security and privacy, and explainability may affect a RS at training and runtime.

Fair Recommender Systems

Despite many EC-proposed provisions regarding AI fairness, in reality, RSs have been shown to provide different recommendation quality to different users, depending on various characteristics, such as gender, age, ethnicity, or personality.[3,7,8,10–12] Such behavior conflicts with the aforementioned goals and is likely to yield unfairness.

Definitions. It is common to distinguish between *individual fairness* and *group fairness*. The former means that similar users are treated in a similar fashion (for example, users with similar skills receive job recommendations within the same pay grade). The latter means that different groups of users defined by some sensitive or protected attribute (for example, gender or ethnicity) are treated in the same way. Accordingly, unfairness is defined as "*systematically and unfairly discriminat[ing] against certain individuals or groups of individuals in favor of others*."[5]

Categories of biases. Unfairness is commonly caused by societal or statistical biases, the former referring to the divergence between how the world should be and how it actually is, the latter to the discrepancy between how the world is and how it is encoded in the system. Such biases can occur at different levels in the recommendation pipeline (see Figure 1). They can be present already in the *data* the algorithms are trained on (for example, an unbalanced dataset with respect to representation of different genders), they can be amplified by the *algorithms* or created *models* (for example, reinforcing stereotypes), or they can originate from users, *cognitive biases* (for example, serial position, anchoring, and decoy effects).[8]

Bias mitigation strategies. To alleviate existing biases, several techniques can be adopted. Focusing on data and algorithm/model bias, the most common approaches are *data rebalancing* (for example, upsampling the minority group of users in the dataset), *regularization* (for example, including a bias correction term in the loss function of the machine/deep learning algorithm), and *adversarial learning* (for example, training a classifier that tries to predict the sensitive attribute from the user-item interaction data and modify the data or recommendation algorithm to minimize the classifier's accuracy).

While much research has been devoted to uncover and mitigate biases in RSs, both within and outside the E.U., many research gaps, the most pressing ones including:

- Several metrics of fairness have been proposed. However, a comprehensive (formal and comparative) study of their strengths and limitations is still missing.[e] Even whether they reflect what humans perceive as fair or unfair—possibly depending on their cultural background, values, and beliefs—has not yet been investigated deeply.

- Likewise, a thorough understanding of capabilities and limitations of existing techniques for mitigating bias through their systematic evaluation is missing.
- From an algorithmic perspective, novel methodologies to debias state-of-the-art RS algorithms, which are predominantly based on deep learning, are needed.
- An investigation of potential economic and social consequences of biases resulting from the use of RSs adopted in high-risk areas (for example, in recruitment) is needed.[3,4]
- Fairness is typically addressed from a system's end user's perspective, but we need to consider multiple RS stakeholders, including content producers, content consumers, and platform providers.
- From a legal perspective, we need provisions with respect to data quality, concrete specifications to whom the obligations to not violate EU non-discrimination law applies, and effective mechanisms for auditing RSs for legal compliance. This requires an interdisciplinary perspective, involving collaboration between researchers with technical expertise and law experts.

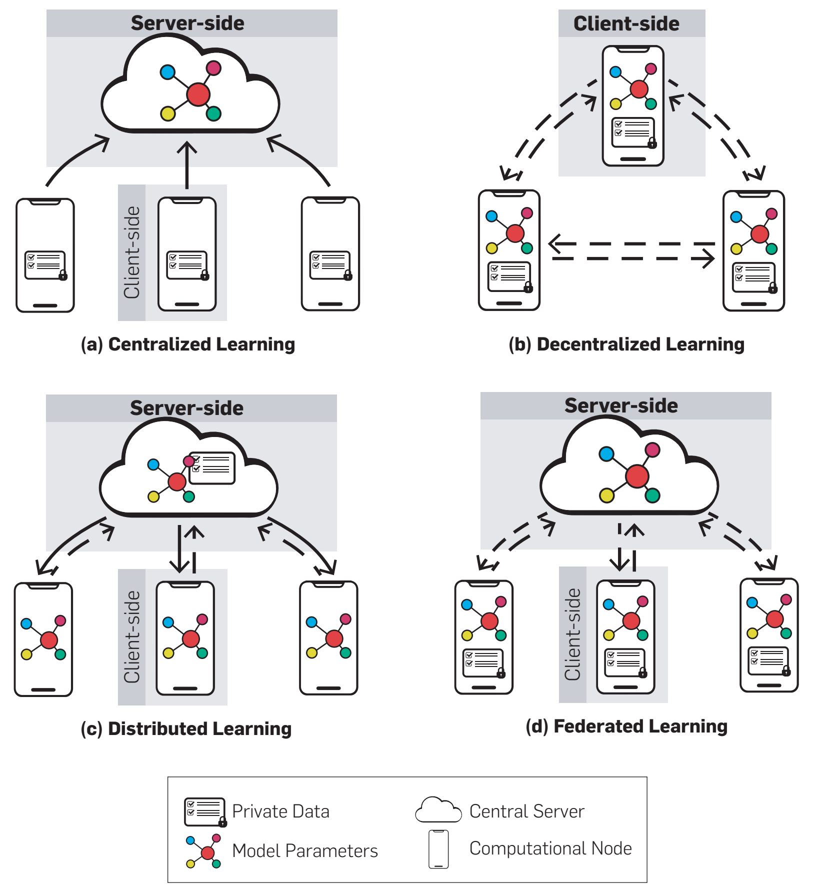
Security and Privacy for Recommender Systems

Privacy in AI is a dominant concern in the EU. To comply with GDPR, AI strategies must be applied considering the new privacy challenges that may limit the uptake of these applications. Privacy-related risks are even more evident when we think about the applications of RSs where user models are

[e] See https://bit.ly/3zPOFZK

Together with attack strategies, defense mechanisms against adversarial attacks have been developed in recent years.

Figure 2. Information flow over the network in four ML architectures. Solid lines represent training data flow, dashed lines represent model parameters flow.



Human oversight is not feasible if explanations are not understandable by people.

built around personal data.

Issues. Data fragmentation and isolation while complying with the GDPR is a major challenge for RS researchers. Actually, preserving users' privacy is not as easy as limiting data collection since a privacy threat may happen at any stage of the data cycle. The model itself stores precious information able to predict future user preferences and behaviors. The main target of privacy attacks in a RS is the confidentiality of the users' sensitive data. Privacy-preserving ML aims to equip ML with defense measures for protecting user privacy and data security. It should be distinguished from secure ML, which attempts instead to preserve integrity and availability of a ML system from intentional (adversarial or poisoning) attacks.

Federated learning. From a privacy perspective, federated learning (FL)[9] completely addresses the principles of focused collection, data minimization, data ownership, and data locality, thus greatly reducing the privacy risks of centralized learning (see Figure 2). While handling users' privacy concerns, FL faces challenges such as communication costs, unbalanced data distribution and device reliability, and security issues (for example, model poisoning, indirect information leakage, and Byzantine adversaries). In Yang et al.,[15] the concept of FL is extended to a more comprehensive idea of privacy-preserving decentralized collaborative ML techniques, both for horizontal federations (where different datasets share the same feature space but are different in training samples) and vertical federations (where different datasets share the training samples but differ in feature space). Thanks to tunable federation approaches in recommendation scenarios, users can be more aware of and decide which data they share.[1]

Unfortunately, while FL can offer significant practical privacy improvements over centralized approaches, there is still no formal guarantee of privacy. This is where other techniques, such as differential privacy, secure multiparty computation, and homomorphic encryption, come to the stage to enforce the privacy protection mechanisms also in recommendation scenarios.

Adversarial attacks and defense. Notwithstanding the great success of machine/deep learning models, recent studies have shown they are not immune to security threats from adversarial use of AI, and the same holds for RSs.[2] An adversary can attack a ML model at two main stages of the learning pipeline, during training or production. These two categories of attacks are respectively known as training-time attack (a.k.a. causative or poisoning attack) and inference-time attack (a.k.a. exploratory or evasion attack).

▸ **Poisoning attack.** Data poisoning attacks are realized by injecting false data points into the training data with the goal to corrupt/degrade the model (for example, the classifier).

▸ **Evasion attack.** Instead of interfering with training data, evasion attacks adjust malicious samples during the inference phase. These attacks are also named *decision-time attacks* referring to their attempt to evade the decision made by the learned model at test time.

Adversarial examples created for image classification tasks are empowered based on continuous real-valued representation of pixels, but in RSs the raw values are user/item identifiers and ratings that are discrete. Hence, adversarial perturbations are added to: the user profile directly (that is, user rating profile), user and item model parameters in a latent factor model; and embeddings representing side information of user and items, respectively.

Together with attack strategies, defense mechanisms against adversarial attacks have been developed in recent years. They can be classified into detection methods and methods seeking to increase the robustness of the learning model. At the heart of the robust optimization method is the assumption that every sample in the training data can be a source for adversarial behavior. It applies a zero-sum game between the prediction and attack adversaries. The ultimate goal in robust optimization is that the prediction model will perform equally well with adversarial and clean inputs.

Explainable Recommender Systems

Although RSs operate as artificial advice givers, people using the system may not understand how the conclusion was reached and when it is appropriate to adopt the advice, or in contrast, when to critique it. The *EU Regulatory Framework for AI* consequently indicates that explanations need to supply human oversight of high-risk systems.

There are three main scientific challenges to overcome before compliance is possible, namely: how to ensure explanations are human-understandable and support appropriate trust; how to build explainable AI (explanation confidence and model complexity); and how to validate the goodness of explanations.

Understandability. Human oversight is not feasible if explanations are not understandable by people, and "*Interpretability*" has been qualified as the degree to which a human can understand the cause of a decision.[13] Understanding is rarely an end-goal in itself, and it is often more useful to measure the effectiveness of explanations in terms of a specific notion of usefulness or explanatory goals such as improved decision support or (appropriate) user trust[14]—minimizing both over- and underreliance on system advice. Furthermore, both the characteristics of the people (for example, expertise, cognitive capacity) and the situation (such as which other people are affected, or which variables are influential) place different requirements on which explanations are useful, also to different presentational choices (for example, with regard to modality, degree of interactivity, level of detail). Simply put: One size does *not* fit all.

Building eXplainable AI (XAI). A large number of methods for XAI have been developed, for a breadth of models and types of data. However, many of them do not (by design) support users in fully understanding the capacities and limitations in a way that would support appropriate trust. We identify two particularly limiting barriers: limited model confidence and high model complexity.

▸ **Confidence.** For sufficient human oversight, RSs must be aware of their knowledge limits not only on the prediction (global and instance) level but also on the explanation level. Consequently, RSs must provide confidence information for each prediction and explanation; and they must clarify how this information has been obtained or computed.

▸ **Complexity.** While it is commonly (but erroneously) believed there is a trade-off between accuracy and interpretability, this is not strictly true. In many cases, several models can offer comparable accuracy performance, but some are more human-understandable. Complexity can be mitigated by selecting the simpler model, and by developing interactive interfaces such as those we have developed in our work, which: adapt the generated explanations to different factors and allow people using the system to see how the factors influence the explanations (transparency), as well as modify the contribution of the factors (control).[6]

Evaluation of explanations. User studies are indispensable for evaluating human performance and understanding. However, to date they are relatively rare in the literature, likely due to their cost. Explanations have also been subjected to automated evaluations, modeled as penalties and constraints in optimization problems, sparsity of the model, monotonicity with respect to a variable, or decomposability into sub-models, and so forth. However, so far there have been no standardized metrics developed. As for ML (Precision, Recall, F-measure, AUC), perhaps this is also because there is no single ideal objective (for example, accuracy versus succinctness). Nevertheless, we hope in the coming years to see benchmarking of such metrics as we see in challenges such as Kaggle, CLEF, and SemEval.

Conclusion

The wide adoption of AI algorithms and systems calls for the definition and realization of a responsible approach to AI. In this respect, by following the documents and legal frameworks proposed over the last years by the EC, some technological issues and trends emerge. We require AI systems to be fair, secure, and privacy-preserving, and interpretable. In this article, we outlined the steps that have already been taken in this direction, as well as indicating what we see as the challenges ahead before we can fulfill the spirit of the European approach to AI. We hope this will serve as a useful roadmap for practitioners and researchers alike.

Acknowledgments. This work has been supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement numbers 860621 and 101031688, by the Linz Institute of Technology, project "Mitigating Gender Bias in Job Recommender Systems: A Machine Learning-Law Synergy" (TIMELY), and by the projects "Casa delle Tecnologie Emergenti della Città di Matera," H2020-ICT-2018-20 "Passepartout," and "Safe and Secure Apulia." T. Di Noia wishes to thank V.W. Anelli, Y. Deldjoo, A. Ferrara, and F. Merra.

References

1. Anelli, V.W. et al. FedeRank: User controlled feedback with federated recommender systems. *ECIR 1* (2021), 32–47.
2. Deldjoo, Y., Di Noia, T., and Merra, F.A. A Survey on adversarial recommender systems: From attack/defense strategies to generative adversarial networks. *ACM Comput. Surv. 54*, 2, Article 35 (Mar. 2022).
3. Fatourou, P., Hankin, C., and Knowles, B. Gender Bias in Automated Decision Making Systems. Policy paper, endorsed by the ACM Europe Technology Policy Committee (2021); https://bit.ly/3r7EHiK
4. Fatourou, P., Papageorgiou, Y., and Petousi, V. Women are needed in STEM: European policies and incentives. *Commun. ACM 62*, 4 (Apr. 2019), 52.
5. Friedman, B. and Nissenbaum, H. Bias in computer systems. *ACM Trans. Inf. Syst. 14*, 3 (July 1996), 330–347.
6. Jin, Y. et al. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction 30*, 2 (2020), 199–249.
7. Lambrecht, A. and Tucker, C.E. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manag. Sci. 65*, 7 (2019), 2966–2981.
8. Lex, E. Psychology-informed recommender systems. *Found. Trends Inf. Retr. 15*, 2 (2021), 134–242.
9. McMahan, B. et al. Communication-efficient learning of deep networks from decentralized data. AISTATS 2017, 1273–1282.
10. Mansoury, M. et al. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM Intern. Conf. Information & Knowledge Management* (2020), 2145–2148.
11. Melchiorre, A.B., Zangerle, E. and Schedl, M. Personality bias of music recommendation algorithms. In *Proceedings of the 14th ACM Conf. Recommender Systems Virtual* (Sept. 2020).
12. Melchiorre, A.B. et al. Investigating gender fairness of recommendation algorithms in the music domain. *Inf. Process. Manag. 58* (2021).
13. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence 267* (2019), 1–38.
14. Tintarev, N. and Masthoff, J. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*. Springer (2015), 353–382.
15. Yang, Q. et al. Federated machine learning: Concept and applications. *ACM TIST 10.2* (2019), 12:1–12:19. doi: 10.1145/3298981.

Tommaso Di Noia is a professor of computer science at Politecnico di Bari, Italy.

Nava Tintarev is a professor of Explainable AI at Maastricht University, and a visiting professor at TU Delft, The Netherlands.

Panagiota Fatourou is a professor of computer science at the University of Crete, Greece.

Markus Schedl is a professor at Johannes Kepler University in Linz, Austria, where he also heads the Human-Centered Artificial Intelligence group at the Linz Institute of Technology AI Lab.